D476–D481 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh125

The Mouse Genome Database (MGD): integrating biology with the genome

Carol J. Bult, Judith A. Blake*, Joel E. Richardson, James A. Kadin, Janan T. Eppig and the Mouse Genome Database Group

The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

Received October 6, 2003; Revised and Accepted October 23, 2003

ABSTRACT

The Mouse Genome Database (MGD) is one component of the Mouse Genome Informatics (MGI) system (http://www.informatics.jax.org), a community database resource for the laboratory mouse. MGD strives to provide a comprehensive knowledgebase about the mouse with experiments and data annotated from both literature and online sources. MGD curates and presents consensus and experimental data representations of genetic, genotype (sequence) and phenotype information including highly detailed reports about genes and gene products. Primary foci of integration are through representations of relationships between genes, sequences and phenotypes. MGD collaborates with other bioinformatics groups to curate a definitive set of information about the laboratory mouse and to build and implement the data and semantic standards that are essential for comparative genome analysis. Recent developments in MGD discussed here include an extensive integration of the mouse sequence data and substantial revisions in the presentation, query and visualization of sequence data.

INTRODUCTION

The Mouse Genome Database (MGD) provides an integrated view of genetic and genomic information for the laboratory mouse (1,2). MGD contains information on mouse genes, genetic markers and genomic features as well as the associations of these features with molecular segments (i.e. probes, primers, cDNA clones, BACs and YACs) and mutant phenotypes. The database also includes comparative mapping data, graphical displays of linkage, cytogenetic and physical maps, experimental mapping data, as well as strain distribution patterns for recombinant inbred strains (RIs), recombinant congenic strains (RCs) and cross haplotypes. MGD is updated daily. A recent snapshot of MGD content is shown in Table 1. Since it first became available on the WWW, MGD has continued to evolve, expanding its data coverage, improving data handling and providing several new data manipulation and display tools.

MGD is one component of the Mouse Genome Informatics (MGI) database resource (http://www.informatics.jax.org) located at the Jackson Laboratory (http://www.jax.org). Other projects and resources that contribute to MGI include the Gene Expression Database (GXD) (3), the Mouse Genome Sequencing (MGS) project and the Mouse Tumor Biology Database (MTB; http://tumor.informatics.jax.org/). The MGI consortium group participates actively in the development and implementation of the Gene Ontologies (GOs) (http://www.geneontology.org). MGI curators also collaborate extensively with Swiss-Prot, RIKEN, and with the LocusLink project at the NCBI to evaluate associations between genes and sequences for the mouse.

IMPROVEMENTS DURING 2003

Integration of sequence data into MGI

A major advance in data representation and integration in the MGI system developed during this year has been to fully integrate mouse sequence and sequence feature annotations. Prior to the integration of sequence data into the database, MGD curators documented the association of sequence data with genes, probes and markers on the basis of curated annotations of literature and/or data loads (e.g. through shared curation of sequence links between MGI and Swiss-Prot, NCBI LocusLink and RIKEN). The connection between sequences and biological entities in the database were represented in MGD only as a table of associations between accession identifiers. Hypertext links were provided from the MGD Gene Detail Pages to the sequence data in GenBank/EMBL/DDBJ and Swiss-Prot. The MouseBLAST (http://mouseblast.informatics.jax.org) server, which is maintained by the MGI consortium, provided a mechanism to query the database using nucleotide or protein sequences. Until recently however, sequence data *per se* were not stored in the MGI database system and users could not query the database using such sequence attributes as strain, sequence type, sequence length, etc. The new enhancements in sequence representation

*To whom correspondence should be addressed. Tel: +1 207 288 6248; Fax: +1 207 288 6132; Email: jblake@informatics.jax.org
The Mouse Genome Database Group: R. M. Baldarelli, K. Barsanti, M. Baya, J. S. Beal, W. J. Boddy, D. W. Bradt, D. L. Burkart, N. E. Butler, J. Campbell, R. Corey, L. E. Corbani, S. Cousins, H. Dene, H. J. Drabkin, K. Frazer, D. M. Garippa, L. H. Glass, C. W. Goldsmith, P. L. Grant, B. L. King, M. Lennon-Pierce, J. Lewis, I. Lu, C. M. Lutz, L. J. Maltais, L. M. McKenzie, D. Miers, D. Modrusan, L. Ni, J. E. Ormsby, D. Qi, S. Ramachandran, T. B. K. Reddy, D. J. Reed, R. Sinclair, D. R. Shaw, C. L. Smith, P. Szauter, B. Taylor, P. Vanden Borre, M. Walker, L. Washburn, I. Witham, J. Winslow and Y. Zhu

Table 1. Snapshot of data content in MGD: September 10, 2003

MGD data statistics	September 10, 2003
Number of references	80436
Number of genes	32417
Number of markers (including genes)	53415
Number of genes with sequence data	29231
Number of markers mapped	49917
Number of mouse/human curated orthologies	9987
Number of genes with links to Swiss-Prot	17214
Number of genes with GO annotations	12527
Number of genes with annotated alleles	3589
Number of annotated alleles	11207
Number of mouse nucleotide sequences curated and integrated in the MGI system (includes ESTs)	>5200000

in the MGI system include the incorporation of all available sequence data for the mouse into MGI, the normalization of sequence feature annotations from NCBI and from providers of large cDNA sequence sets, and the development of query and visualization components specific to sequence representations.

The sequence data that have been and are being integrated into the MGI database system include the RIKEN full-length cDNA clone data sets (4), the Mammalian Gene Collection (MGC) clone data (5), computational gene models from Ensembl and the NCBI's annotation of the reference mouse genome sequence (6), and 'virtual' transcript data sets from The Institute for Genomic Research (TIGR Mouse Gene Index) and the Computational Biology and Informatics Laboratory (CBIL) at the University of Pennsylvania (7). Sequence and clone information for such commonly used clone libraries as the National Institute on Ageing (NIA) set and the RPCI-23 and -24 BAC clones have also been incorporated into the database. For each sequence set, a combination of manual and computational approaches for sequence to gene correlations assured the correct data associations [see (7–9) for detailed descriptions of sequence annotation processes]. Each sequence–gene association is documented independently so that modifications in the gene–sequence sets can been facilitated. MGI curation staff normalized sequence feature annotations including (i) strain names, (ii) library name, (iii) type of sequence and (iv) tissue of origin. Sequence features represented in MGI include these four attributes as well as sequence length, reference of origin and clone collection (if applicable). As a result, users can submit complex queries based on sequence attributes and filtered by other gene-specific information such as functional attributes or time/tissue of expression (Fig. 1). Sequence summary reports (Fig. 2) and sequence detail reports (Fig. 3) incorporate detailed information about the sequence. It is now possible for users of the MGI system to download sequence data directly from the database instead of having to follow a series of hypertext links before they can access the actual sequences associated with a gene or other marker in MGD.

Enhanced queries by gene

A new feature on the 'Genes and Markers Query Form' provides an improved option for gene name text searches. Whereas previous nomenclature queries examined current symbols/names and synonyms, the new default query also searches the allele symbols/names and human gene symbols/names. The summary results present a list of matches on the query term ordered by matches on the current symbol, current name, allele symbol, allele name, withdrawn symbol, withdrawn name, synonym, human ortholog symbol, human ortholog name and all other ortholog symbols. Each item of information is linked to a detail page. If the information being sought is not present on the list, it is possible to determine, at a glance, how to narrow the query by altering the search string.

Enhanced gene detail reports

Two major enhancements of the gene detail page include the presentation of short statements about the phenotype(s) associated with the gene, and a new presentation of summary information about the sequence of the gene (Fig. 4). The short phenotype descriptions are tied to the manual curation of model mutant and allele phenotype data and thus will be reviewed each time new mutant phenotype data is annotated for the gene. The summary sequence data expands the representation of map position to include the sequence coordinates (currently from the EnsEMBL annotation of NCBI Build 30) and a link to the EnsEMBL ContigView and to the UCSC Browser. The incorporation of links from MGI to the MapViewer at NCBI are planned.

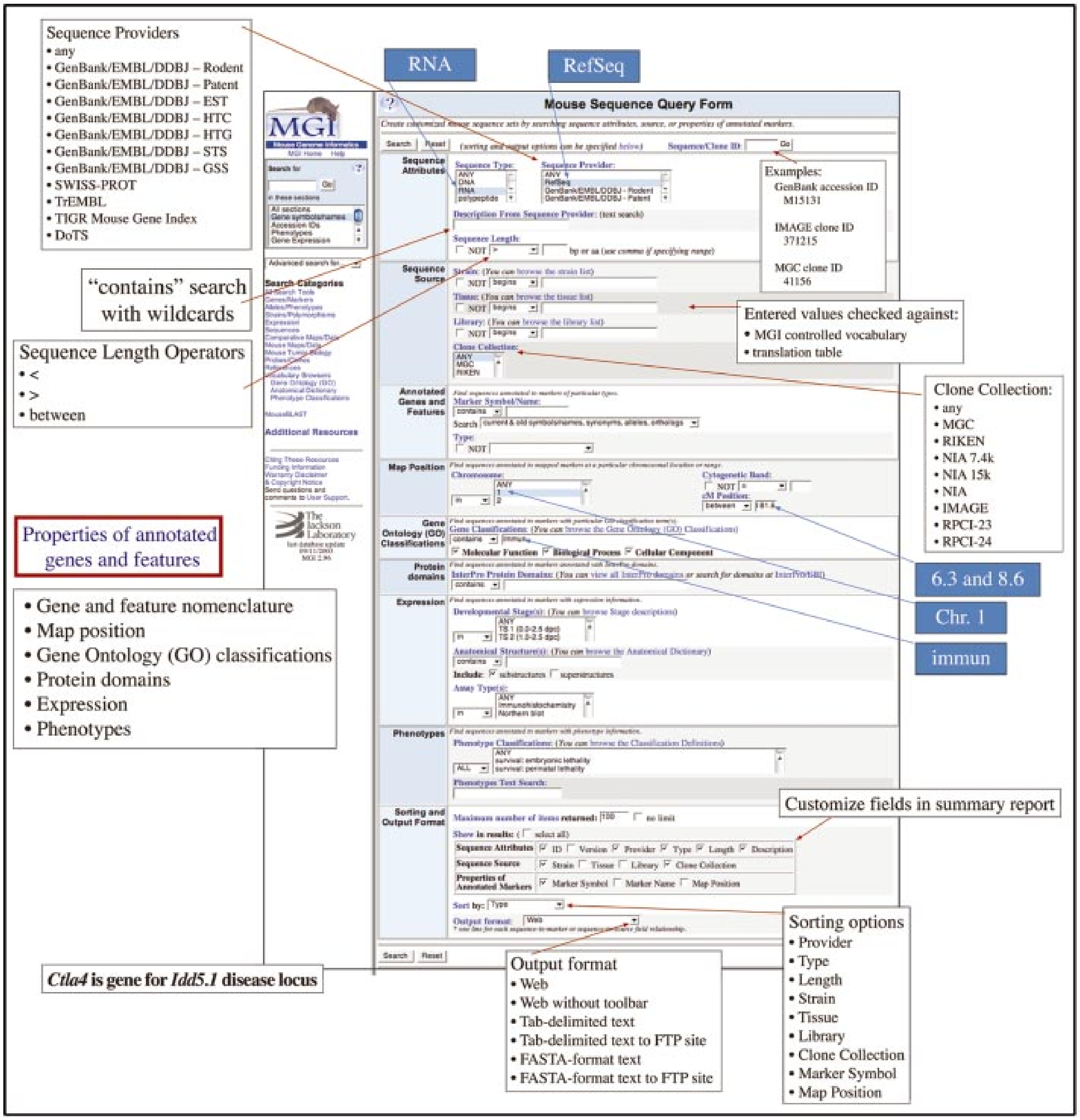
OTHER INFORMATION

User input

MGD encourages user input into its gene and allele annotation efforts. On each gene detail and allele detail page, a clickable button ('Your Input Welcome') brings the user to a web-based form for submitting updates to the information being viewed.

Mouse gene nomenclature

The MGD gene annotation group assigns unique symbols and names to mouse genes under the guidelines set by the International Committee on Standardized Genetic Nomenclature for mouse (http://www.informatics.jax.org/mgihome/nomen/index.shtml). Through curation of shared links between MGI and other bioinformatics resources, the official nomenclature for mouse genes is becoming widely disseminated. The MGI nomenclature group works closely



Sequence Providers
- any
- GenBank/EMBL/DDBJ – Rodent
- GenBank/EMBL/DDBJ – Patent
- GenBank/EMBL/DDBJ – EST
- GenBank/EMBL/DDBJ – HTC
- GenBank/EMBL/DDBJ – HTG
- GenBank/EMBL/DDBJ – STS
- GenBank/EMBL/DDBJ – GSS
- SWISS-PROT
- TrEMBL
- TIGR Mouse Gene Index
- DoTS

"contains" search with wildcards

Sequence Length Operators
- <
- >
- between

Properties of annotated genes and features

- Gene and feature nomenclature
- Map position
- Gene Ontology (GO) classifications
- Protein domains
- Expression
- Phenotypes

Ctla4 is gene for *Idd5.1* disease locus

RNA

RefSeq

Mouse Sequence Query Form

Examples:
GenBank accession ID M15131
IMAGE clone ID 371215
MGC clone ID 41156

Entered values checked against:
- MGI controlled vocabulary
- translation table

Clone Collection:
- any
- MGC
- RIKEN
- NIA 7.4k
- NIA 15k
- NIA
- IMAGE
- RPCI-23
- RPCI-24

6.3 and 8.6

Chr. 1

immun

Customize fields in summary report

Sorting options
- Provider
- Type
- Length
- Strain
- Tissue
- Library
- Clone Collection
- Marker Symbol
- Map Position

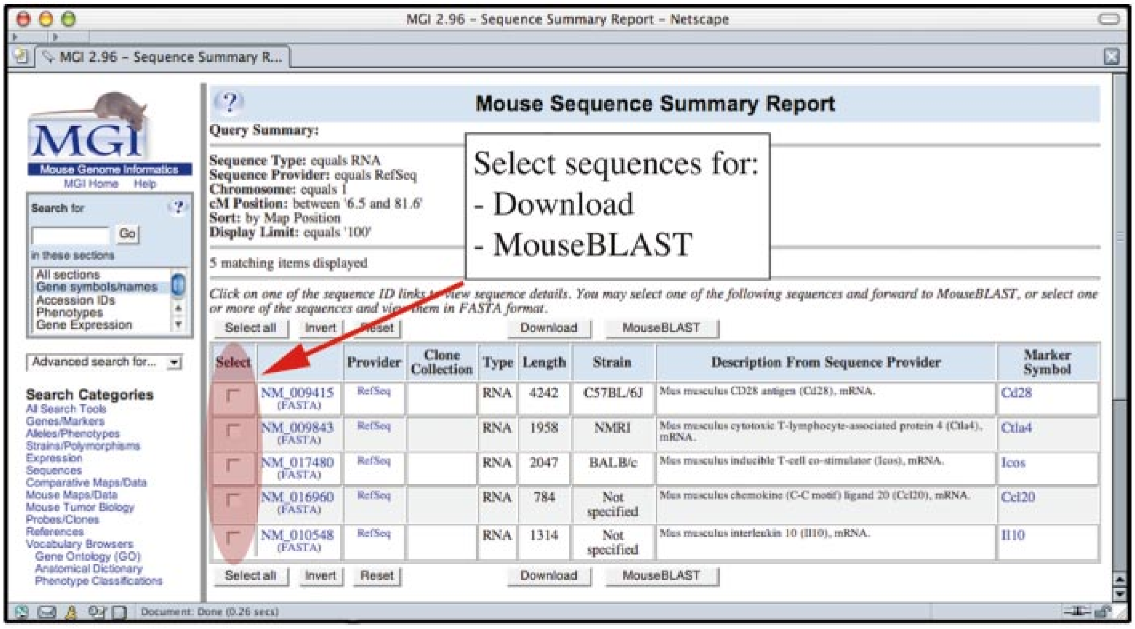
Output format
- Web
- Web without toolbar
- Tab-delimited text
- Tab-delimited text to FTP site
- FASTA-format text
- FASTA-format text to FTP site

Figure 1. Mouse Sequence Query Form. The new Mouse Sequence Query Form allows the user to query by sequence attributes, source attributes and properties of annotated markers including: gene symbol/name, map position, GO classification, expression (anatomical structure and Theiler stage) and phenotype. One application of the new sequence query form is positional candidate analysis, where users can request sequences annotated to genes with properties of interest within a chromosomal interval. In this example, this query form can be used to find candidates for the Idd5.1 (insulin-dependent Diabetes Mellitus 5) quantitative trait locus (QTL) that have gene products involved in the immune response and/or have defense/immunity protein activity. The query form has been set to find mRNAs for these genes that could then be used to design PCR primers for expression analysis. The query searched for RefSeq mRNAs annotated to genes within the Chr.1 interval from 6.3 to 8.6 cM with GO terms containing 'immun'.

with nomenclature specialists for human (http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl) and rat (http://rgd.mcw.edu) to provide consistent nomenclature for mammalian species. Scientists can reserve symbols prior to publication using the electronic nomenclature submission form (http://www.informatics.jax.org/mgihome/nomen/nomen_submit_form.shtml) or by contacting the MGD nomenclature coordinator by email (nomen@informatics.jax.org). The MGD nomenclature coordinator can also assist with other nomenclature issues such as revision of gene family designations.

Electronic data submission

Any type of data that MGD maintains can be submitted as an electronic contribution, although mapping data, polymorphisms and mammalian homologies are currently the most common. Each electronic submission receives a permanent database accession ID. All data sets are associated with either an electronic submission reference or a published paper. MGD reference pages provide links to associated data sets. Online information about data submissions procedures is found at http://www.informatics.jax.org/mgihome/.



Mouse Sequence Summary Report

Query Summary:

Sequence Type: equals RNA
Sequence Provider: equals RefSeq
Chromosome: equals 1
cM Position: between '6.5 and 81.6'
Sort: by Map Position
Display Limit: equals '100'

5 matching items displayed

Click on one of the sequence ID links to view sequence details. You may select one of the following sequences and forward to MouseBLAST, or select one or more of the sequences and view them in FASTA format.

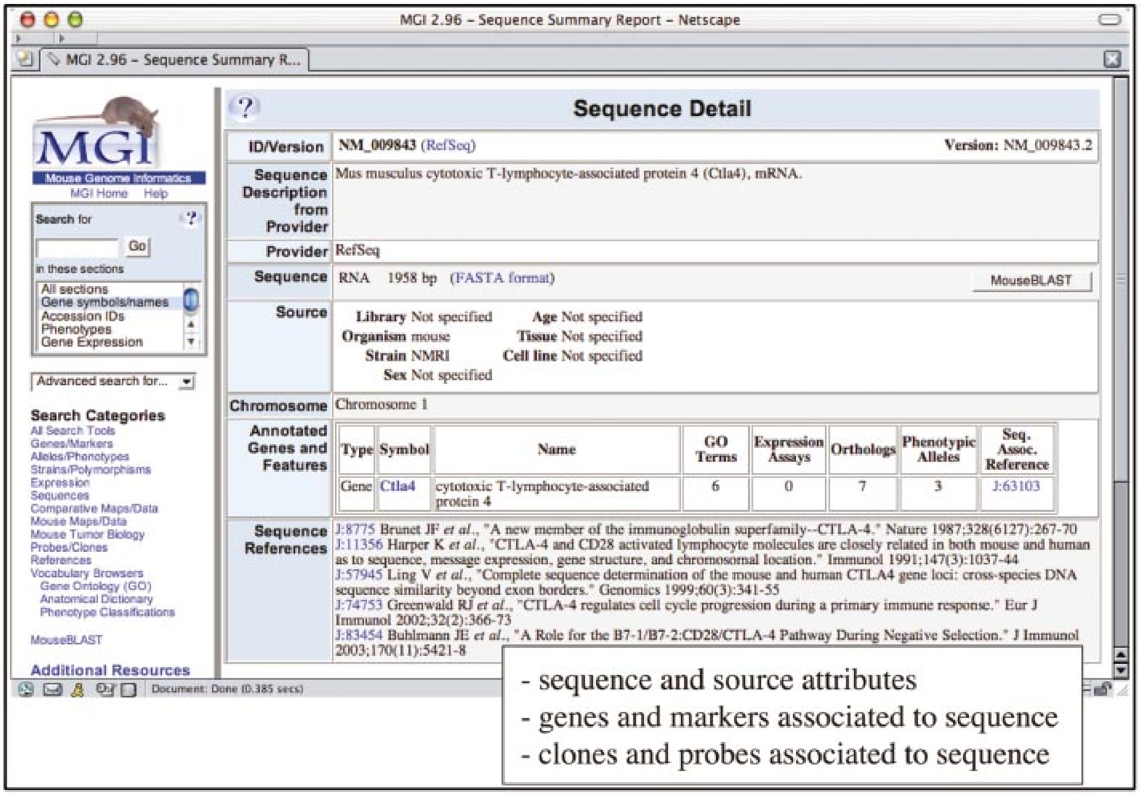
Select all | Invert | Reset | Download | MouseBLAST

Select		Provider	Clone Collection	Type	Length	Strain	Description From Sequence Provider	Marker Symbol
□	NM_009415 (FASTA)	RefSeq		RNA	4242	C57BL/6J	Mus musculus CD28 antigen (Cd28), mRNA.	Cd28
□	NM_009843 (FASTA)	RefSeq		RNA	1958	NMRI	Mus musculus cytotoxic T-lymphocyte-associated protein 4 (Ctla4), mRNA.	Ctla4
□	NM_017480 (FASTA)	RefSeq		RNA	2047	BALB/c	Mus musculus inducible T-cell co-stimulator (Icos), mRNA.	Icos
□	NM_016960 (FASTA)	RefSeq		RNA	784	Not specified	Mus musculus chemokine (C-C motif) ligand 20 (Ccl20), mRNA.	Ccl20
□	NM_010548 (FASTA)	RefSeq		RNA	1314	Not specified	Mus musculus interleukin 10 (Il10), mRNA.	Il10

Select all | Invert | Reset | Download | MouseBLAST

Select sequences for:
- Download
- MouseBLAST

Figure 2. Sequence Summary Report. The expanded Sequence Summary Report displays the query summary at the top of the page, and provides details about the sequences returned by the query including the type, length, strain and description. In addition to links to a Sequence Detail Report, sequences can be selected and either downloaded in FASTA format or forwarded to MouseBLAST. This Summary Report shows the results from the query described in Figure 1 and shows a mRNA for each of the five genes that matched the query criteria. These genes are *Cd28*, *Ctla4*, *Icos*, *Ccl20* and *Il10*.



Sequence Detail

ID/Version	NM_009843 (RefSeq) Version: NM_009843.2
Sequence Description from Provider	Mus musculus cytotoxic T-lymphocyte-associated protein 4 (Ctla4), mRNA.
Provider	RefSeq
Sequence	RNA 1958 bp (FASTA format) MouseBLAST
Source	**Library** Not specified **Age** Not specified **Organism** mouse **Tissue** Not specified **Strain** NMRI **Cell line** Not specified **Sex** Not specified
Chromosome	Chromosome 1

Annotated Genes and Features

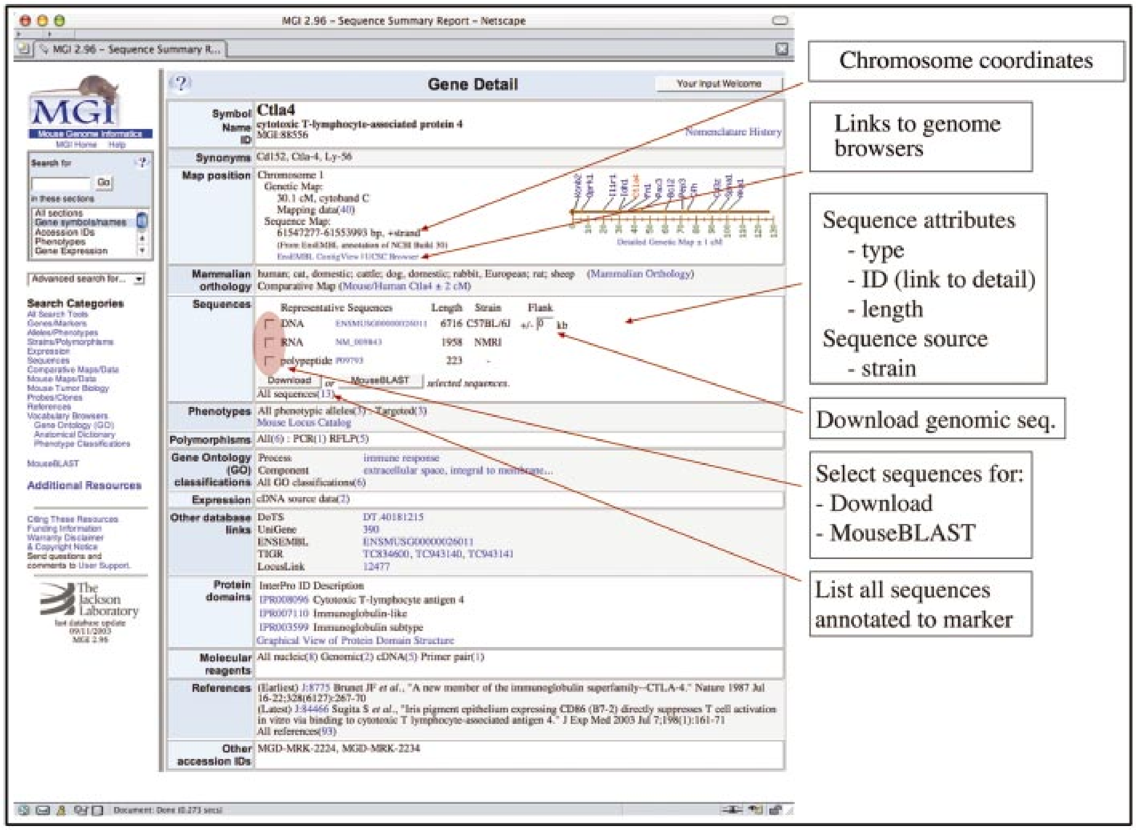
Type	Symbol	Name	GO Terms	Expression Assays	Orthologs	Phenotypic Alleles	Seq. Assoc. Reference
Gene	Ctla4	cytotoxic T-lymphocyte-associated protein 4	6	0	7	3	J:63103

Sequence References

J:8775 Brunet JF *et al*., "A new member of the immunoglobulin superfamily--CTLA-4." Nature 1987;328(6127):267-70
J:11356 Harper K *et al*., "CTLA-4 and CD28 activated lymphocyte molecules are closely related in both mouse and human as to sequence, message expression, gene structure, and chromosomal location." Immunol 1991;147(3):1037-44
J:57945 Ling V *et al*., "Complete sequence determination of the mouse and human CTLA4 gene loci: cross-species DNA sequence similarity beyond exon borders." Genomics 1999;60(3):341-55
J:74753 Greenwald RJ *et al*., "CTLA-4 regulates cell cycle progression during a primary immune response." Eur J Immunol 2002;32(2):366-73
J:83454 Buhlmann JE *et al*., "A Role for the B7-1/B7-2:CD28/CTLA-4 Pathway During Negative Selection." J Immunol 2003;170(11):5421-8

- sequence and source attributes
- genes and markers associated to sequence
- clones and probes associated to sequence

Figure 3. Sequence Detail report. The new Sequence Detail report for the RefSeq record, NM_009843, shows the sequence attributes, source attributes, and the genes that have been annotated to this sequence. A summary of additional data curated for the gene is displayed with links to the detail reports for those data.



Gene Detail

Symbol Ctla4
Name cytotoxic T-lymphocyte-associated protein 4
ID MGI:88556

Synonyms Cd152, Ctla-4, Ly-56

Map position Chromosome 1
Genetic Map:
30.1 cM, cytoband C
Mapping data(40)
Sequence Map:
61547277-61553093 bp, +strand
(From EnsEMBL annotation of NCBI Build 30)
EnsEMBL ContigView | UCSC Browser

Mammalian orthology human; cat, domestic; cattle; dog, domestic; rabbit, European; rat; sheep (Mammalian Orthology)
Comparative Map (Mouse/Human Ctla4 ± 2 cM)

Sequences

	Representative Sequences		Length	Strain	Flank
□	DNA	ENSMUSG00000026011	6716	C57BL/6J	+/- 0 kb
□	RNA	NM_009843	1958	NMRI	
□	polypeptide	P09793	223	-	

Download or MouseBLAST selected sequences.
All sequences(13)

Phenotypes All phenotypic alleles(3) - Targeted(3)
Mouse Locus Catalog

Polymorphisms All(6) : PCR(1) RFLP(5)

Gene Ontology (GO) classifications
Process immune response
Component extracellular space, integral to membrane...
All GO classifications(6)

Expression cDNA source data(2)

Other database links
DoTS DT.40181215
UniGene 390
ENSEMBL ENSMUSG00000026011
TIGR TC834600, TC943140, TC943141
LocusLink 12477

Protein domains InterPro ID Description
IPR008096 Cytotoxic T-lymphocyte antigen 4
IPR007110 Immunoglobulin-like
IPR003599 Immunoglobulin subtype
Graphical View of Protein Domain Structure

Molecular reagents All nucleic(8) Genomic(2) cDNA(5) Primer pair(1)

References (Earliest) J:8775 Brunet JF *et al.*, "A new member of the immunoglobulin superfamily--CTLA-4." Nature 1987 Jul 16-22;328(6127):267-70
(Latest) J:84466 Sugita S *et al.*, "Iris pigment epithelium expressing CD86 (B7-2) directly suppresses T cell activation in vitro via binding to cytotoxic T lymphocyte-associated antigen 4." J Exp Med 2003 Jul 7;198(1):161-71
All references(93)

Other accession IDs MGD-MRK-2224, MGD-MRK-2234

Chromosome coordinates

Links to genome browsers

Sequence attributes
- type
- ID (link to detail)
- length
Sequence source
- strain

Download genomic seq.

Select sequences for:
- Download
- MouseBLAST

List all sequences annotated to marker

Figure 4. Gene Detail report. The expanded Gene Detail report now shows the coordinates for the gene on the NCBI Build 30 mouse genome sequence, and displays representative DNA (genomic), RNA and polypeptide sequences. The genomic sequence for the gene, plus a selectable amount of flanking sequence, can be downloaded in FASTA format or forwarded to MouseBLAST. Here a Gene Detail Report for the gene *Ctla4* is displayed.

Community outreach and user support

MGD provides extensive user support through online documentation and easy email or phone access to User Support Staff: User Support WWW access: http://www.informatics.jax.org/mgihome/support/support.shtml; email access: mgi-help@informatics.jax.org; telephone access: +1 207 288 6445; fax access: +1 207 288 6132.

Other outreach. MGI-LIST (http://www.informatics.jax.org/mgihome/lists/lists.shtml), is a moderated and active email bulletin board supported by the MGI Users Support group. Other outreach includes online tutorials and answers to frequently asked questions, available at: http://www.informatics.jax.org/userdocs/helpdocs_menu.shtml. Lee Silver's book, *Mouse Genetics*, is now available in an electronic version at http://www.informatics.jax.org/silver/. The online version has been enhanced by linking genes and references to MGI and MEDLINE.

IMPLEMENTATION

MGD is implemented in the Sybase relational database system, version 12.5. A large set of CGI scripts and Java Servlets mediates the user's interaction with the database. For computational users, direct SQL access can be requested through User Support. User-requested database reports and a number of widely used data files (generated daily) are available on the FTP site (ftp://ftp.informatics.jax.org).

CITING MGD

The following citation format is suggested when referring to data sets specific to the MGD component of MGI: Mouse Genome Database (MGD), Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine (URL: http://www.informatics.jax.org). [Type in date (month, year) when you retrieved the data cited.] For general citation of the Mouse Genome Informatics (MGI) resource please cite this article.

ACKNOWLEDGEMENTS

MGD is supported by NIH/NHGRI grant HG00330.

REFERENCES

1. Blake,J.A., Richardson.J.E., Bult,C.J., Kadin,J.A., Eppig,J.T. and the Mouse Genome Database Group (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.*, **31**, 193–195.
2. Blake,J.A., Eppig,J.T., Richardson,J.E., Bult,C.J., Kadin,J.A. and the Mouse Genome Database Group (2002) The Mouse Genome Database (MGD): the model organism database for the laboratory mouse. *Nucleic Acids Res.*, **30**, 113–115.
3. Hill,D.P., Begley,D.A., Finger,J.H., Hayamizu,T.F., McCright,I.J., Smith,C.M., Beal,J.S., Corbani,L.E., Blake,J.A. *et al.* (2004) The mouse

gene expression database (GXD): updates and enhancements. *Nucleic Acids Res.*, **32**, D568–D571.

4. FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
5. Strausberg,R.L., Feingold,E.A., Klausner,R.D. and Collins, F.S. (1999) The mammalian gene collection. *Science*, **286**, 455–457.
6. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* Mouse Genome Sequencing Consortium and Mouse Genome Analysis Group (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
7. Zhu,Y., King,B.L., Parvizi,B., Brunk,B.P., Stoeckert,C.J.,Jr, Quackenbush,J., Richardson,J., and Bult,C.J. (2003) Integrating computationally assembled mouse transcript sequences with the Mouse Genome Informatics (MGI) database. *Genome Biol.*, **4**, R16.
8. Baldarelli,R.M., Hill,D.P., Blake,J.A., Adachi,J., Furuno,M., Bradt,D., Corbani,L.E., Cousins,S., Frazer,K.S., Qi,D. *et al.* (2003) Connecting sequence and biology in the laboratory mouse. *Genome Res.*, **13**, 1505–1519.
9. Kasukawa,T., Furuno,M., Nikaido,I., Bono,H., Hume,D.A., Bult,C., Hill,D.P., Baldarelli,R., Gough,J., Kanapin,A. *et al.* (2003) Development and evaluation of an automated annotation pipeline and cDNA annotation system. *Genome Res.*, **13**, 1542–1551.